DOCUMENT RESUME

ED 344 932 TM 018 281

AUTHOR	Littlefield, John; And Others
TITLE	Analyzing Written Comments by Performance Raters.
PUB DATE	Apr 92
NOTE	7p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Behavior Patterns; *Classification; Evaluation Methods; *Evaluators; Examiners; Higher Education; *Interrater Reliability; Medical Education; *Medical Students; *Rating Scales; Surgery
IDENTIFIERS	*Letters of Recommendation; *Performance Based Evaluation

ABSTRACT

A four-level taxonomy is proposed to define the usefulness of rater written comments for supporting letters of recommendation. The taxonomy is used to classify comments on 220 rating forms by 25 raters from two surgery departments regarding performance by third-year medical students. Written comments were classified by the following taxonomy: (1) no comment; (2) vague comment (e.g., overall good performance); (3) descriptive comment (e.g., excellent rapport with patients); and (4) behaviorally referenced comment (e.g., excellent rapport with patients as indicated by numerous complimentary comments from patients). Behaviorally referenced comments were deemed the most useful to directors of clerkships. Two reviewers were trained to read and classify the written comments, and interrater reliability was assessed by having the raters read and classify 20 of the same forms. Results indicate that individual raters within a department and departments at separate sites differ in the percentage of written rater comments at each of the four levels. While individual rater differences account for differences in comment types, perceived departmental expectations also appear to influence the levels of written rater comments. There is a 13-item list of references. (SLD)

Analyzing Written Comments by Performace R .ters

John Littlefield, Debra DaRosa, Richard Bell, and Gary Nicholas

Paper presented at the Annual Meeting of the American Educational Research Association

San Francisco, California, April 20, 1992

Analyzing Written Comments by Performance Raters

John Littlefield, Debra DaRosa, Richard Bell, and Gary Nicholas

Background

Performance ratings are the most widely used method for evaluating student proficiency in professions education. In a typical medical education setting, raters observe students over a one to eight week period and then assign a numerical score plus write narrative comments to explain their rationale for the score. The scores for a given student are combined across multiple raters to generate a mean score for grade assignment. The various written comments are synthesized by the clerkship director into a written paragraph describing the student's strengths and weaknesses. Paragraphs from multiple clerkships are eventually used to write a recommendation letter when the student applies for postgraduate residency training.

Under ideal conditions, the various numerical rating scores for a given medical student will be quite similar and the written comments will provide behaviorally-referenced documentation of the student's performance. Under actual conditions, inter-rater reliability is in the .25 - .35 range (Maxim & Dielman, 1987) and written comments by many raters are marginally useful for constructing a recommendation letter. This study analyzes written comments by individual medical faculty rating third-year students from the perspective of how useful the comments are in providing support for recommendation letters.

In the 1960's and 70's, most performance rating research focused on characteristics of the rating form. The underlying assumption seemed to be that a better definition of the criteria for making judgments would improve the reliability and validity of the resulting scores. In a comprehensive review of the performance ratings research, Landy and Farr (1980) recommended a moratorium on rating form research. They proposed that researchers look more closely at how raters process information to make judgments.

Rater information processing was a major theme of performance ratings research in the 1980's. For example, Nathan and Alexander (1985) describe two general roles for raters: observer-recorder or evaluator-judge. The observer-recorder's cognitive task entails fastidious record keeping for later use by a decision maker. By contrast, the evaluator-judge must draw conclusions about a student's performance typically using a "global rating." Cadwell and Jenkins (1986) found that teachers' implicit theories about student behavior influence their ratings of hypothetical students. Cadwell and Jenkins describe the "rater as the measuring instrument," reflecting a profound shift in the focus of performance ratings research from the rating form to rater perception, information processing, and judgment. These insights regarding different rater roles and rater implicit theories about students help clarify rater information processing. However, they do not provide practical guidelines to improve the numerical accuracy of ratings for assigning grades or the adequacy of comments for writing recommendation letters.

The most direct route for improving the evaluative usefulness of performance ratings data is to improve the skills of individual raters. The skill levels of individual raters in making numerical ratings has been shown to be quite variable. Marienfeld & Reid (1984) documented the existence of overly lenient raters during a six year period. Littlefield et. al. (1991) defined individual rater accuracy as stringency or leniency in relation to all other raters who evaluated the same cohort of students. The proportion of "accurate raters" varied from 26% to 75% across five surgery departments. In summary, previous research has provided empirically-based definitions of rater skill in providing accurate numerical ratings, but no corresponding definitions are available for assessing rater skills in writing narrative comments. This study proposes a four level taxonomy that defines the usefulness of rater written comments for supporting letters of recommendation.

The taxonomy is used to classify comments on 220 rating forms by 25 raters from two surgery departments.

Methods

Written comments by 25 raters regarding performance by third-year medical students in two surgery departments served as data for this study. Department One provided 111 forms completed by 14 raters (45% of 247 forms completed during 1988-89) while Department Two provided 211 forms completed by 11 raters. A stratified random sample of 109 Department Two forms (52%) was selected to reduce the labor burden (stratified by beginning, middle, and end of the academic year). All raters who completed five or more performance ratings were included in the sampled data.

Written comments by performance raters were classified on a four level taxonomy: 1. no comment, 2. vague comment (e.g., overall good performance), 3. descriptive comment (e.g., excellent rapport with patients), 4. behaviorally-referenced comment (e.g., excellent rapport with patients as indicated by numerous complimentary comments from patients). Behaviorally-referenced comment^ were deemed most useful to clerkship directors for three reasons: to document the rationale for numerical ratings, to provide quotes for use in recommendation letters, and to provide constructive feedback to students.

Two reviewers were trained to read and classify written comments into one of the four categories. If multiple comments were noted on a single form, the form was coded according to the highest level comment. Inter-rater agreement was analyzed by having both reviewers classify twenty rating forms. The two reviewers agreed on 93% of their classification decisions after adjusting down for chance agreements (Tinsley & Weiss, 1975). Each reviewer then read and classified comments from 1/2 of the remaining 200 rating forms used in the study.

Data analyses addressed three research questions:

1. Do individual raters within a department differ in the level of their written comments?
2. Do the two departments differ in the level of their written comments?
3. Can individual raters be classified as writing comments at predominately one taxonomy level?

Research questions 1 and 2 above were addressed by calculating Kruskal-Wallis one-way analysis of variance by ranks (Siegel, 1956). This nonparametric statistic tests whether differences among the sample scores (raters or departments) signify genuine population differences or whether they represent chance variations. The null hypothesis is that the raters (departments) all come from the same population. Research question 3 was addressed by calculating the proportion of comments by each rater in each of the four categories. If a rater wrote 50% or more comments in a given category then she/he was classified as predominately at that level (e.g., descriptive comments).

Results

Table 1 displays the tabulated results from the reviewers' classification of the written comments. A Kruskal-Wallis test of rater scores from Department One revealed significant differences among individual raters ($X^2 = 52.41$, df = 13, $p \le .001$). These differences can be observed by noting that raters 1, 6, and 10 never wrote comments while raters 7 and 9 wrote predominately Descriptive comments. Rater scores from Department 2 were also tested and differences among individual raters were statistically significant ($X^2 = 37.51$, df = 10, $p \le .001$). Raters in this department typically wrote Descriptive comments (70%) although rater 7 wrote 67% Behaviorally-referenced comments while rater 8 wrote 50% Vague comments. Differences in the

quality of written comments between the two departments were also statistically significant ($X^2 = 80.87$, $df = 1$, $p \leq .001$). In response to research question 3, 24 of the 25 raters in the study wrote 50% or more of their comments at one taxonomy level. The only exception was rater 11 from department 1.

Table 1 - Classification of Rater Written Comments

	None	Vague	Descriptive	Beh.-referenced
Department 1				
Rater #				
1	5	0	0	0
2	1	6	1	0
3	1	0	3	2
4	1	0	3	2
5	0	5	0	0
6	7	0	0	0
7	0	2	5	0
8	13	0	0	1
9	1	2	6	0
10	5	0	0	0
11	2	2	1	0
12	5	1	4	0
13	5	1	4	0
14	12	0	3	0
Total - Dept. 1	58	19	29	5
Department 2				
Rater #				
1	0	0	13	0
2	0	2	8	2
3	0	0	8	2
4	0	1	6	5
5	0	1	11	0
6	0	0	3	1
7	0	0	4	8
8	1	5	4	0
9	0	2	9	1
10	0	0	7	0
11	0	0	3	2
Total - Dept. 2	1	11	76	21

Discussion

It appears that individual raters within a department and departments at separate sites differ in the percentage of written rater comments at each of the four taxonomic levels. It also appears that individual raters can be classified as writing comments predominately at one of the four taxonomic levels therefore this behavior appears to be relatively stable over time. Differences among individual raters could be attributed to motivation levels. These raters are busy surgery faculty and in the absence of positive reinforcement, writing a vague comment or nothing at all may be a means of saving time. Differences in the taxonomic levels of written comments between the two

departments could be attributed to perceived expectations among raters. It appears that Department Two raters expect that rating forms should not be submitted without written comments (1% without comments) while Department One raters did not perceive this expectation (52% without comments). These interpretations regarding rater motivation and perceived expectations seem plausible but do not provide useful guidelines to clerkship directors who would like to increase the frequency of behaviorally-referenced comments for use in writing recommendation letters.

Another interpretation of these results is to view the rating forms as one link in a communication system designed to help raters convey their observations numerically and narratively to the clerkship director. The cumulative communication regarding a given student is successful if the clerkship director has reliable numerical data to assign a grade and sufficient written comments to construct an insightful paragraph for a recommendation letter. For example, the system could be functioning satisfactorily even though there are no written comments on the rating forms provided the clerkship director has sufficient verbal communication regarding individual student performance (e.g., a small close-knit department). However, if the clerkship director frequently has disparate numerical ratings and insufficient narrative comments then this study's general research approach of analyzing individual rater behavior offers several guidelines for improving rater skills, the most important link in the communication system.

Rater skills, viewed from a human learning perspective, can be defined as a *cognitive contextual module*, a unified complex of knowledge, skills, goals, and feelings of an individual in relation to some activity (Bereiter, 1990). Over time these separate cognitive components integrate into a coherent whole (i.e., module). A given rater's contextual module regarding the task of evaluating student clinical performance consists of knowledge regarding her role (e.g., observer-recorder or evaluator-judge), personal interest in teaching responsibilities (e.g., implicit theories regarding students), personal experiences with ratings received as a medical student, and perceptions of colleagues' attitudes and motivation toward their role as raters. This hypothesized cognitive contextual module would be activated whenever the rater is asked to evaluate a student.

If a rater's cognitive module is highly developed, he will typically produce accurate numerical ratings (Littlefield, et. al., 1991) and narrative comments that are useful to the clerkship director in writing a paragraph for a recommendation letter. If the rater's cognitive module needs to be further developed, the clerkship director has several options. For a given cohort of students (e.g., an academic year), the rater could receive a numerical summary of his ratings and those by colleagues who rated the same students plus the clerkship director's summary narrative paragraphs for each student in the cohort group. The rater could also go on rounds with another rater whose cognitive module is highly developed and discuss the performance levels of the students he observed. These actions are a form of rater training, but hopefully the rater will perceive them as uniquely designed to help him communicate more effectively with the clerkship director.

Conclusions

This study analyzes individual rater behavior as a means to improve the communication of rater observations to clerkship directors. Most raters write 50% or more of their comments at one of the four taxonomic levels. Some individual performance raters routinely write comments at a higher taxonomic level than their colleagues. This finding regarding individual rater skill differences in writing narrative comments parallels reports that some raters are better calibrated numerically than others (Littlefield, et. al.,1991). Differences in the taxonomic levels of written comments were also observed at the department level. This finding suggests that perceived departmental expectations also influence the levels of written rater comments.

Future performance ratings research in medical education should continue this approach of analyzing individual rater behavior. The next step is to understand the *local meanings* of both

numerical ratings and narrative comments from the rater's and clerkship director's point of view (Erickson, 1986). Over time, a clerkship director probably develops implicit expectations regarding numerical ratings and written comments from a given rater. Certain key words may prompt the director to personally contact the rater for a more in-depth explanation.

Performance ratings research focused on medical faculty numerical and narrative interpretations of student behavior was first published thirty-three years ago (Cowles and Kubany, 1959). This early rating form was developed through a critical incident technique and provided detailed instructions and specimen comments describing effective and ineffective behavior for each of the eight student characteristics. In a follow-up study, Cowles (1965) analyzed 2300 rater comments related to each of the eight student characteristics and rated them on a six point scale of goodness of student performance. More recently, Rhoton (1989) described a performance rating system based entirely upon written comments. Rhoton's research and this study could be viewed as a rebirth of the recognition that both numerical ratings and narrative comments are integral parts of a communication link between raters and clerkship directors. Future research should focus on better understanding how the various elements of the communication system (students, raters, numerical and narrative data, and clerkship directors) interact with one another.

Bibliography

Bereiter, C. (1990). Aspects of an educational learning theory. *Rev. of Educ. Res.*, V. 60, 603-624.

Cadwell, J. & Jenkins, J. (1986). Teachers' judgments about their students: the effect of cognitive simplification strategies on the rating process, *Am. Educ. Res. J.*, V. 23, 460-475.

Cowles, J. (1965). A critical-comments approach to the rating of medical students' clinical performance, *Journal of Medical Education*, V40(2), 188-198.

Cowles, J., & Kubany , A. (1959). Improving the measurement of clinical performance of medical students. *J. Clin. Psychol.*, V 15, 139-142.

Erickson, F. (1986). Qualitative methods in research on teaching, in M.C. Wittrock (Ed.) *Handbook on Research on Teaching*, (3rd ed., pp. 119-161). New York: Macmillan.

Landy, F. & Farr, J. (1980). Performance rating. *Psych. Bull.*, V 87, 72-107.

Littlefield, J., DaRosa, D., Anderson, K., Bell, R. Nicholas, G. & Wolfson, P. (1991). Accuracy of surgery clerkship performance raters, *Academic Medicine*, V 66(9), supplement, S16-S18.

Marienfeld, R.D. & Reid, J.C. (1984). Six-year documentation of the easy grader in the medical clerkship setting, *Journal of Medical Education*, V 59, 589-591.

Maxim, B. & Dielman, T. (1987). Dimensionality, internal consistency and interrater reliability of clinical performance ratings, *Medical Education*, 21, 130-137.

Nathan, B. & Alexander, R. (1985). The role of inferential accuracy in performance rating, *Academy of Management Review*, V 10(1), 109-115.

Rhoton, M. (1989). A new method to evaluate clinical performance and critical incidents in anaesthesia: Quantification of daily comments by teachers, *Medical Education*, 23, 280-289.

Siegel. S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Tinsley, H. & Weiss, D. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Couns. Psych.*, V. 22, 358-376.